Published in final edited form as:

J Proteome Res. 2019 March 01; 18(3): 814–825. doi:10.1021/acs.jproteome.8b00875.

Identification of Hybrid Insulin Peptides (HIPs) in Mouse and Human Islets by Mass Spectrometry

T. Aaron Wiles[†], Roger Powell[†], Cole Robert Michel[†], K. Scott Beard[‡], Anita Hohenstein^{†,§}, Brenda Bradley[§], Nichole Reisdorph[†], Kathryn Haskins[§], Thomas Delong^{*,†}

[†]Department of Pharmaceutical Sciences, Skaggs School of Pharmacy and Pharmaceutical Sciences

[§]Department of Immunology and Microbiology, School of Medicine, University of Colorado Anschutz Medical Campus, Aurora, Colorado 80045, United States

[‡]Barbara Davis Center for Childhood Diabetes, Aurora, Colorado 80045, United States

Abstract

We recently discovered hybrid insulin peptides (HIPs) as a novel class of post-translationally modified peptides in murine-derived beta cell tumors, and we demonstrated that these molecules are autoantigens in type 1 diabetes (T1D). A HIP consists of an insulin fragment linked to another secretory granule peptide via a peptide bond. We verified that autoreactive CD4 T cells in both mouse and human autoimmune diabetes recognize these modified peptides. Here, we use mass spectrometric analyses to confirm the presence of HIPs in both mouse and human pancreatic islets. We also present criteria for the confident identification of these peptides. This work supports the hypothesis that HIPs are autoantigens in human T1D and provides a foundation for future efforts to interrogate this previously unknown component of the beta cell proteome.

Graphical Abstract

*Corresponding Author thomas.delong@ucdenver.edu. Tel: 303-724-0448.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the [ACS Publications website](https://pubs.acs.org/doi/10.1021/acs.jproteome.8b00875) at DOI: 10.1021/acs.jproteome.8b00875. Summary of peptides identified in BALB/c mouse islets by mass spectrometry. False discovery rates (FDRs) for database searches. Validation of putative HIPs identified in the islets of BALB/c mice. Additional mirror plots demonstrating validation of HIPs identified in NOD islets. Putative HIPs identified by mass spectrometry in islets from BALB/c mice. Additional mirror plots demonstrating validation of HIPs. Pearson correlation coefficients (PCC). (PDF)

The authors declare the following competing financial interest(s): A provisional patent application has been submitted for the use of hybrid insulin peptides and HIP-reactive T cells as biomarkers in T1D (TD, KH).

Keywords

pancreatic islets; beta cell proteome; mass spectrometry; type 1 diabetes (T1D); hybrid insulin peptide (HIP)

INTRODUCTION

Using mass spectrometric analysis, we previously identified a new class of post-translationally modified peptides present in murine pancreatic beta cell tumors: hybrid insulin peptides (HIPs).^{1,2} HIPs consist of insulin peptides that have been linked to other peptides via a traditional peptide bond to generate new amino acid sequences that are not encoded in the genome (Scheme 1). We and others demonstrated that HIPs are autoantigens for pathogenic CD4 T cells in the human autoimmune disease type 1 diabetes (T1D)^{1,3} and in the nonobese diabetic (NOD) mouse model of the disease.^{1,2,4,5} In T1D, destruction of the insulin-producing beta cells of the pancreas by the immune system results in chronic hyperglycemia. Much effort has been focused on identifying the beta cell proteins being recognized by the aberrant immune response in T1D, and post-translational modification of proteins has been proposed as a mechanism by which autoantigens can be generated from germline-encoded proteins.^{6,7} The confident identification of HIPs is a critical step in the study of the role of these modified peptides as autoantigens in T1D and will expand our characterization of the beta cell proteome.

Mass spectrometry is currently the best available technology for identifying peptides and proteins—particularly those with novel sequences—in complex biological samples in a high-throughput fashion. In standard mass spectrometry-based proteomics workflows, tandem mass spectra are searched against a database of known, genomically encoded protein sequences. Because traditional post-translational modifications (PTMs) such as citrullination and oxidation simply introduce a known mass shift at a specific amino acid residue within a

protein, consideration of such modifications can be easily integrated into standard workflows. HIPs, in contrast, represent a unique and more complicated category of modification that cannot be identified by searching normal databases using predefined mass shifts as modifications. Unlike conventional PTMs, HIP formation leads to the substitution of an entire region of amino acid sequence with a different sequence. This modification can affect a large number of amino acid residues and is not restricted to a single chemical change and a predictable mass shift. These challenges cannot be circumvented simply by searching custom databases containing all possible HIP sequences, as the size of such databases would lead to prohibitively long search times and an increased risk of false discoveries. The latter concern is exacerbated by the fact that, when using the entire proteome as a template, many short peptide regions with similar but distinct amino acid sequences can be found. Thus, such databases could contain many hybrid sequences that closely resemble genomically encoded sequences, increasing the likelihood of false HIP matches.

The confident identification of HIPs is further complicated by a challenge that is increasingly apparent in the peptidomics field. In traditional bottom-up mass spectrometry-based proteomics workflows, the presence of a specific protein in a given sample is generally determined through the identification of several different peptides originating from proteolytic digestion of that protein. Because different proteotypic peptides are identified, less confidence is needed in the assignment of any one peptide identification. However, identification of a HIP requires that the peptide found spans the distinctive hybrid peptide junction; therefore, for a given digest, identification will often be based on interpretation of a single MS/MS spectrum. As has been recently discussed,⁸ such peptide-centric, or “peptidomics”, workflows require that individual spectra be analyzed more carefully and present a new set of challenges to the proteomics field.

As the study of HIPs in the field of T1D research gains momentum and reports on the identification of potential hybrid peptide sequences begin to emerge,^{9,10} there is a pressing need to establish rigorous guidelines to confidently identify new HIPs and other types of hybrid peptides. Here, we introduce a mass spectrometry-based workflow that makes use of a focused HIP database and a rigorous, multilayered validation approach to confidently identify HIPs in biological samples while reducing the risk of false positive discoveries. Using this workflow, we confirm the presence of HIPs in primary mouse islets and for the first time demonstrate that HIPs are also present in human islets. The work presented here highlights the underappreciated complexity of the beta cell proteome and the need for specialized approaches to characterize proteins that are missed by conventional proteomics workflows.

MATERIALS AND METHODS

Islet Isolation

Isolation of mouse islets was conducted under a protocol approved by the Institutional Animal Care and Use Committee. BALB/c mice or nondiabetic NOD mice were anesthetized and the pancreas was inflated via the common bile duct with collagenase solution and then removed. Following incubation at 37 °C to facilitate digestion, islets were isolated by density centrifugation. Islets were then handpicked under a microscope and

aliquoted into microcentrifuge tubes containing medium supplemented with fetal bovine serum (5%). Tubes were spun at 300g for 5 min, supernatant was removed, and the tubes were placed in liquid nitrogen to rapidly freeze the islet pellets. Islets were stored at -80 °C until use. Because individual mice only yield approximately 150–250 islets, islets were harvested from 5 to 8 mice of the same strain at a time and islets from all of the mice were pooled. Replicate experiments were performed using islets from different harvests. Islets isolated from deceased nondiabetic human donors (a 36-year-old female, a 44-year-old male, and a 36-year-old female) were obtained from Prodo Laboratories (Aliso Viejo, CA, USA) and the University of Alberta Diabetes Institute Islet Core (Edmonton, Alberta, Canada) and were likewise frozen as pellets without medium until use. Each human islet experiment was performed using islets from a different donor. A small number of islets (400 islet equivalents for mouse experiments and 500 for human experiments) were used as starting material for each experiment. To account for the varying size of individual islets, islet quantities are reported as the number of islet equivalents rather than the number of islets. One islet equivalent corresponds to a standardized area (size) when counting islets under a microscope; a single islet can correspond to more or less than one islet equivalent, depending upon its size.

Sample Preparation

For mouse and human experiments, 400 or 500 islet equivalents, respectively, were thawed, resuspended in 50 μL of 1 $\times$ phosphate-buffered saline (PBS) and 50 μL of trifluoroethanol (TFE), and sonicated for 5 min. Islets were then heated at 95 °C for 10 min, vortexing every 2 min, and then sonicated for an additional 5 min. Samples were spun at 17,000g for 5 min to pellet debris. The supernatant was separated isocratically on a Waters 7.8 mm $\times$ 150 mm XBridge BEH size exclusion chromatography (SEC) column (3.5 μm particles, 200 Å pores) with a 7.8 mm $\times$ 30 mm guard column using 25 mM ammonium acetate (pH $\approx$ 7) as the running buffer at a flow rate of 1 mL/min for 10 min. Four fractions of equal volume were collected between 3.8 and 7.4 min. A 100 μL aliquot of each fraction was used to determine protein concentration by a bicinchoninic acid assay (Pierce) and the remainder of each fraction was dried using a vacuum concentrator. Dried fractions were reconstituted with 50 μL of 50 mM ammonium bicarbonate (pH 8.0) and sonicated for 5 min. For mouse islets, the protease AspN (Promega) was then added to yield a final enzyme:protein ratio of 1:20 (w/w) and zinc sulfate was added as an enzyme cofactor to a final concentration of 0.5 mM. The protease AspN was selected because its specificity was best suited to generate peptide fragments from HIPs containing insulin C-peptide based on the mouse C-peptide sequence. Human islet samples were digested with GluC (Thermo), which was more suitable than AspN for digestion of the human insulin C-peptide sequence, at a final enzyme:protein ratio of 1:20 (w/w). Reactions were incubated overnight at 37 °C and then dried in a vacuum concentrator. Samples were reconstituted in loading buffer (3% acetonitrile/0.1% formic acid/water), sonicated, and spun at 17,000g for 2 min to remove any insoluble material. The supernatant was then analyzed by liquid chromatography tandem-mass spectrometry (LC-MS/MS).

LC-MS/MS Analysis

Digested SEC fractions were analyzed by LC-MS/MS using an Agilent 1200 series UHPLC system with a nanoflow adapter and an Agilent 6550 Q-TOF equipped with a nano-ESI source. Online sample separation was accomplished by reversed-phase liquid chromatography using a Thermo Acclaim Pepmap 100 C18 trap column (75 $\mu\text{m} \times 2 \text{ cm}$; 3 μm particles; 100 Å pores) and Thermo Acclaim Pepmap RSLC C18 analytical column (75 μm inner diameter; 2 μm particles; 100 Å pores) in a trap forward-elute configuration using a water/acetonitrile gradient (buffer A: 0.1% formic acid in water; buffer B: 0.1% formic acid and 90% acetonitrile in water). Mass spectrometry data was collected in positive ion mode with an MS scan range of 290–1700 m/z and acquisition rate of 5 spectra/s and an MS/MS scan range of 50–1700 m/z and a minimum scan rate of 3 spectra/s. Abundance dependent accumulation, which varies the MS/MS scan speed based on precursor abundance, was enabled with a target of 40,000 counts/spectrum. Ten precursors were automatically selected for fragmentation per cycle based on abundance, excluding singly charged precursors, with an absolute threshold of 3000 counts and a relative threshold of 0.01%. After the first run of each sample, the data were searched against the standard SwissProt database using the Spectrum Mill MS Proteomics Workbench (Agilent, Rev B.06.00.201) and the sample was run again excluding precursors that were confidently identified in the first analysis in order to improve coverage of lower abundance precursors. The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium¹¹ via the PRIDE¹² partner repository with the data set identifier PXD011606. A README.txt file is also provided as a key to the file names.

Initial Spectrum Mill Analysis

Data were analyzed using the Spectrum Mill software. During extraction, MS/MS scans within a 2 min window were merged based on precursor selection purity, spectral similarity, retention time, and m/z . Merged spectra were searched against the SwissProt mouse or human database with Agilent ESI Q-TOF defined as the instrument. The precursor mass tolerance was set at ± 10 ppm and the product ion mass tolerance was set at ± 20 ppm. No enzymatic digest was defined to allow for identification of naturally truncated peptides, which we and others have shown to be present in beta cells.^{1,9,13,14} Matches were then validated with the following set of thresholds: score > 10 , percentage scored peak intensity (SPI) $> 70\%$, and rank 1 minus rank 2 (R1-R2) score > 2.5 .^{15–18} SPI indicates the percentage of the total ion intensity in the MS/MS spectrum that can be explained by the sequence match. R1-R2 score indicates the difference between the scores for the best and second best sequence match found in the database in a particular search and provides an additional assessment of confidence in a match. Spectra that did not validate were searched against a custom mouse or human database containing hypothetical HIPs (see below) using the same mass tolerances and validation thresholds as used in the first search. As in the initial search, no enzymatic digest was specified. To determine if any of the spectra that were matched to a HIP sequence could be reasonably explained by a nonhybrid sequence containing one or more oxidized methionine residues, these spectra were researched against the SwissProt database using the same parameters as in the initial search but allowing methionine oxidation as a potential modification. The precursor mass shift range was set at 0 to 100 Da to allow for oxidation of multiple methionine residues in a single precursor.

For both SwissProt and custom database searches, spectra were concurrently searched against a reverse sequence database to enable determination of false discovery rates (FDRs). Spectrum Mill calculates FDRs by multiplying the number of validated hits in the reverse sequence database (“false” or “decoy” hits) by two and then dividing the result by the total number of hits across both databases (“true” and “false” hits) and multiplying by 100 to generate a percentage. Doubling of the reverse database hits is based on the assumption that the number of false validated hits in the forward database search is equal to the number of validated hits in the reverse database search.¹⁹ However, an assessment of this approach concluded that it overestimates the actual FDR.²⁰ Thus, we calculated FDRs manually using the traditional approach of dividing the number of validated hits in the reverse sequence (decoy) database hits by the number of validated hits in the forward (target) database and multiplying by 100.²⁰ If Spectrum Mill matches multiple spectra to the same peptide sequence, the multiple matches can be grouped and only the spectrum match with the highest score will be reported as the representative distinct peptide match. Reported FDRs are based on the number of distinct peptide matches as reported by Spectrum Mill rather than matches at the spectral level.

HIP Sequence Databases

An in-house computer algorithm was used to generate a 3000-entry mouse HIP database and a 1271-entry human HIP database containing each possible C-terminal truncation of insulin C-peptide linked to every predicted naturally occurring cleavage product of insulin, chromogranin A (ChgA), islet amyloid polypeptide (IAPP), secretogranin 1 (Scg1), and neuropeptide Y (NPY). For the mouse database, sequences of both insulin isoforms, insulin 1 and insulin 2, were included. Because the IAPP sequence in the NOD strain differs from the sequence in other mice,^{2,21,22} both variants were included to allow searching of data from various strains. The HIP sequence databases were concatenated with the traditional SwissProt databases so that, when calculating R1-R2 scores, the Spectrum Mill software would consider matches to traditional nonhybrid sequences. These custom databases are available for download as part of the ProteomeXchange Consortium¹¹ data set PXD011606 as described above.

Validation of Putative HIPs

To assess the validity of putative HIP matches in greater detail, several additional criteria were applied. These criteria are discussed in detail in the results section. When determining the b- and y-ion coverage of the left and right peptide region of each potential HIP, only b- and y-ions considered for scoring by the Spectrum Mill software were counted. For scoring, the software considered the 25 most abundant peaks remaining in a given fragmentation spectrum after peak filtering. Thus, additional b- and y-ions may have been present but were not counted due to low abundance. Singly and doubly charged ions were counted, but not ions corresponding to neutral losses.

Synthetic peptides used for validation were obtained commercially from Synpeptide (Shanghai, China) or CHI Scientific (Maynard, MA, USA) at >70% purity. Identical acquisition methods were used for mass spectrometric analysis of islet and synthetic peptide samples. Fragmentation spectra of endogenous islet peptides identified as potential HIPs and

the corresponding synthetic peptide spectra were compared using the Spectrum Mill Spectrum Matcher tool to assess similarity. The 25 most abundant ^{12}C peaks were considered for scoring, with a precursor m/z tolerance of ± 10 ppm and a product ion m/z tolerance of ± 20 ppm. Additionally, spectral similarity was assessed by calculating the Pearson correlation coefficient (PCC). To calculate the PCC, first the mean and standard deviation of the peak intensities were determined for the endogenous and synthetic peptide spectra separately. An in-house algorithm was then used to pair peaks from the two spectra that differed by less than 20 ppm; each peak was only counted once. Peaks that only existed in one spectrum were assigned an abundance of zero in the other spectrum. Peaks that were not present at an intensity greater than three standard deviations above the mean in at least one of the two spectra were considered noise and were removed. The PCC of the processed data set was then calculated using GraphPad Prism software.

For retention time comparison, synthetic peptides were analyzed using the same instrument configuration and method used for the islet sample. Both the islet sample and synthetic peptide were spiked with the PROCAL peptide retention time calibration mixture (JPT Peptide Technologies, Germany)²³ and observed retention times were adjusted based on the retention times of the calibrants to account for run-to-run variability and matrix effects. A point-to-point linear model was used for retention time alignment, in which only two calibrants—the one eluting immediately before and the one eluting immediately after the peptide of interest—were considered, similar to the “on-the-fly” approach used by Escher et al.²⁴ In early experiments, retention time standards were not used and thus for some of the reported HIPs, accurate retention time comparisons could not be made.

RESULTS

Preparation of Islet Proteins for Mass Spectrometric Analysis

We previously used mass spectrometry to identify two different HIPs that are targeted by autoreactive CD4 T cells in NOD mice: 2.5HIP¹ and 6.9HIP.^{1,2} 2.5HIP contains the peptide sequence DLQTLAL-WSRM (hyphen indicates the hybrid peptide junction) and results from the fusion of an insulin C-peptide fragment and the chromogranin A (ChgA) peptide WE14. 6.9HIP contains the sequence DLQTLAL-NAAR and results from the fusion of the same C-peptide sequence and a propeptide region of islet amyloid polypeptide (IAPP). We refer to the peptide that becomes the N-terminal portion of the HIP as the left peptide and the peptide that becomes the C-terminal portion as the right peptide. Because isolation of mouse islets is very laborious and yields only a small amount of tissue, beta cell tumors from NOD.RIP-TAg mice²⁵ were used as the starting material in these experiments. Based on this original protocol, we developed a microscale sample preparation procedure for preparing frozen primary islets for subsequent identification of HIPs by mass spectrometry. Four- or five-hundred frozen islet equivalents (mouse or human origin) were suspended, heated, and sonicated in 50% trifluoroethanol (TFE) to lyse cells and liberate proteins. The organic solvent TFE is regularly used to denature proteins and can be easily removed from samples using a vacuum concentrator. This makes it more compatible with downstream mass spectrometric analysis than detergents, leading potentially to improved sensitivity compared to detergent-based sample preparation approaches.²⁶ Following lysis, cellular debris was

removed by centrifugation and the remaining proteins and peptides were separated by size exclusion chromatography (SEC). Although SEC is a low resolution technique not often used in protein/peptide discovery, it is effective at separating the smaller molecules of interest in our experiments, such as insulin and HIPs, from larger proteins, such as enzymes and structural proteins. Fractions were collected in the lower molecular weight range and were dried using a vacuum concentrator.

Chromatographic fractions were next reconstituted and digested to generate peptides for mass spectrometric analysis. Trypsin, which cleaves peptides at the C-terminus of arginine and lysine residues, is the most commonly used protease in bottom-up proteomics workflows because it tends to generate multiply charged peptides that are of suitable size for mass spectrometric analysis. Secretory granule proteins contain dibasic residue sites (any two amino acid combination of lysine and/or arginine residues) that are cleaved naturally by processing enzymes during the conversion of the precursor versions of these proteins to the mature, fully processed forms. However, since there tends to be a paucity of lysine and arginine residues in the resulting peptides, trypsin was not well suited for this application. We therefore digested fractions with either the enzyme AspN (mouse samples) or GluC (human samples), which cleave at the N-terminal side of aspartic acid (Asp, D) residues and the C-terminal side of glutamic acid (Glu, E) residues, respectively. These proteases were chosen to yield peptides of optimal length for various potential mouse and human HIPs. For the present investigation, we focused on identifying HIPs containing C-peptide fragments on the left and a limited number of possible peptides on the right, few of which contain cysteine residues. Therefore, the samples were not reduced and alkylated, simplifying the workflow.

LC-MS/MS and Initial Data Analysis

Following digestion, the samples were analyzed by LC-MS/MS using a nanoscale C18 reversed-phase column on an ultrahigh performance liquid chromatography (UHPLC) system and an Agilent 6550 quadrupole time-of-flight (Q-TOF) mass spectrometer equipped with a nanoelectrospray ionization (nano-ESI) source. Fragmentation was carried out by collision-induced dissociation (CID), using N₂ as collision gas. In order to improve coverage of low abundance peptides, each sample was run twice, with precursors confidently identified in the first run being excluded from selection for fragmentation in the second run. MS/MS data were first searched against the SwissProt mouse proteome database using the Spectrum Mill software (Agilent) to identify traditional genomically encoded peptides.

Spectra that could not be confidently assigned to known peptides were then searched against a custom database (Scheme 2) containing hypothetical HIPs. To reduce search times and false discoveries, we used a focused HIP database rather than attempting to account for all potential HIP combinations. The 2.5HIP and 6.9HIP previously identified by mass spectrometry each consist of an insulin C-peptide fragment on the N-terminal side, and we consistently observe numerous C-terminally truncated C-peptide fragments by mass spectrometry when analyzing beta cell tumors or islets. Therefore, we used all possible C-terminal truncations of insulin C-peptide as the left peptides in the database. The 2.5HIP and 6.9HIP each consist of a naturally occurring cleavage product of another beta cell secretory granule protein as the right peptide. Consequently, we limited the possible right peptides in

our database to predicted natural cleavage products of major beta cell secretory granule proteins. This yielded a 3000-entry mouse HIP database and a 1271-entry human HIP database.

Summaries of all the peptides (HIPs and non-HIPs) identified in islets from NOD mice (Figure 1a), BALB/c mice (Figure S-1), and human donors (Figure 1b) by this two-stage recursive workflow are provided. HIPs constituted less than 2% of the total peptides identified, with 46 HIPs being tentatively identified in NOD islets samples, 28 in BALB/c islet samples, and 30 in islet samples from human donors. For both the SwissProt and HIP database searches, spectra were also searched against reverse sequence databases and false discovery rates (FDRs) were calculated. The average FDR at the distinct peptide level for all searches and all samples was 0.48% for NOD islets, 0.56% for BALB/c islets, and 0.92% for human islets (Figure S-2).

Validation of HIPs

To minimize the possibility of incorrectly assigning interpretations to MS/MS spectra, we established a list of seven criteria, or rules, for validating hybrid peptides (Table 1). The first five rules relate to information obtained directly from the MS/MS spectrum of the peptide identified in the islet sample (endogenous peptide). The final two rules involve comparisons between information collected about the endogenous peptide and a synthetic version of the proposed HIP sequence. The application of these criteria to validate putative HIPs found in the islets of NOD mice and nondiabetic human donors is demonstrated in Table 2 and Table 3, respectively. Results of the validation of HIP matches obtained upon analysis of BALB/c islets are provided in Table S-1. Only those matches that satisfied the first three criteria are included in the tables. For each match, values that met a particular criterion are highlighted in green and those that did not meet the criterion are highlighted in yellow. Matches that met all criteria are highlighted in green in the “Sequence” column of the table. Although the rigor of these criteria may exclude some true matches, the objective was to eliminate all but those matches that were identified with highest confidence.

The first two rules were applied automatically using standard capabilities of the Spectrum Mill software. First, any spectra that could be confidently matched to a traditional nonhybrid peptide represented in the SwissProt database were not searched against the HIP database. Three parameters were used for automated filtering of results within the Spectrum Mill software: percentage scored peak intensity (%SPI, which indicates the percentage of the total intensity of scored peaks that can be explained by the proposed sequence match), score, and rank 1 minus rank 2 (R1-R2 score). R1-R2 score reports the difference between the score for the best match and the score for the second best match. A low R1-R2 score indicates that the observed spectrum can be described by two different sequences with similar confidence. In other words, the spectrum is ambiguous.

When searching for HIPs, it is also important to account for the possibility that a spectrum matched to a HIP sequence may be better described by a nonhybrid sequence containing a traditional post-translational modification. Thus, after completion of the initial SwissProt search and the HIP database search, we performed a second search of only those spectra matched to HIP sequences (see Table 2, Table 3, and Table S-1) against the SwissProt

database, allowing methionine oxidation as a potential modification. Methionine oxidation was selected due its prevalence as a physiological modification and as an artifact of sample preparation.²⁷ The results of this search were used to update the results of the previous searches and are reflected in the database search results provided in the tables. None of the spectra matched to HIPs could be confidently explained by a nonhybrid sequence containing oxidized methionine.

In Table 2, Table 3, and Table S-1, the %SPI, score, and R1-R2 score are reported for the best match found for each spectrum in the traditional SwissProt database (including matches to sequences containing oxidized methionine), demonstrating that the spectrum did not confidently match to a traditional peptide sequence (Rule 1). For example, as shown in Table 3, an MS/MS spectrum acquired by analysis of human islets was matched to the putative HIP sequence LGGGPGAGSLQPL-EAE. In the initial search of the standard SwissProt database, the best match for this spectrum had a % SPI of 46.9, a score of 8.1, and a R1-R2 score of 0.8, indicating that the spectrum could not be confidently matched to a nonhybrid sequence. The HIP match was then required to surpass the same stringent %SPI, score, and R1-R2 score thresholds (Rule 2). The best match for the spectrum in this search was the proposed HIP sequence LGGGPGAGSLQPL-EAE. This match surpassed all three requirements with a %SPI of 100.0 ($100.0 > 70$), a score of 17.0 ($17.0 > 10$), and a R1-R2 score of 4.5 ($4.5 > 2.5$). Because hypothetical HIP sequences often differ from corresponding native peptide sequences by only a few amino acids, the same spectrum can sometimes be matched with similar confidence to both a native and a hybrid sequence. The R1-R2 score provides a valuable metric for identifying such ambiguous spectra and was thus a critical factor in determining the validity of HIP matches.

Islet samples contain many naturally-truncated peptides,^{1,9,13,14} and no digest was specified for either the SwissProt database search or the HIP database search in order to allow for identification of these peptides. However, as an additional safeguard to prevent false identification of HIPs, only those potential HIP matches that corresponded to the proteolytic digest used in sample preparation were chosen for further consideration (Rule 3). The sequence LGGGPGAGSLQPL-EAE, for example, would be naturally preceded by a glutamic acid residue and contains a glutamic acid residue at its C-terminus and would thus correspond to a digest with the enzyme GluC. Although one missed cleavage following a glutamic acid residue exists within this sequence, such missed cleavages are common and are allowed in most searches. A sample could certainly contain HIP fragments that would not be predicted based on the specificity of the protease used during sample preparation. However, application of this requirement adds another layer of confidence to the filtered matches. By manually applying this criterion after the searches but allowing for nonspecific cleavages in the searches, naturally truncated peptides were able to contribute to R1-R2 scores.

When matching a spectrum to a nonhybrid sequence, the presence of fragment ions that specifically describe only one-half of the proposed sequence may still be sufficient to engender confidence in the match. However, matching a spectrum to a HIP sequence indicates that the full sequence contains two different regions originating from two different proteins or noncontiguous regions of the same protein. Thus, it is important to confirm that

the observed MS/MS spectrum provides sufficient evidence to identify both the left and right peptide sequences. First, we required that the HIP sequence identified contained at least two left peptide residues and two right peptide residues (Rule 4). As the number of residues in each side of the HIP increases, the protein from which a given side originates can be assigned with greater confidence. The human HIP sequence LGGPGAGSLQPL-EAE was designated as an insulin C-peptide/C-peptide HIP (see Table 3). Because the left peptide contains 13 residues, this sequence can be confidently and uniquely matched to insulin C-peptide. (It should be noted that the mass spectrometric techniques used cannot distinguish between leucine (L) and isoleucine (I) residues; for some peptides, this ambiguity could complicate the determination of the protein of origin). Because the right peptide EAE is only three residues in length, we cannot confidently assign it to a distinct protein of origin. However, given the abundance of insulin in islets, this sequence most likely originated from insulin C-peptide. Next, at least two band/or y-ions (singly- and/or doubly-charged) specifically describing the left peptide sequence and at least two describing the right peptide sequence were required (Rule 5). A b- or y-ion corresponding to fragmentation between the left and right peptide (i.e., fragmentation at the HIP junction) was counted toward this criterion for both the left and right peptide b/y-ion coverage. For the spectrum matched to the sequence LGGPGAGSLQPL-EAE, five ions describing the left peptide and three ions describing the right peptide were identified, giving us confidence that the proposed HIP sequence accurately describes the endogenous peptide.

Validation of HIPs: Comparison to Synthetic Peptide Standards

As a more rigorous assessment of the validity of each match, synthetic peptides corresponding to the proposed HIP sequences were obtained and analyzed by LC-MS/MS. The chromatographic retention time and MS/MS spectrum for the synthetic peptide were compared to those for the endogenous islet peptide to determine similarity. Spectral similarity was determined objectively using Spectrum Matcher, a tool available in the Spectrum Mill software, and numerical results are provided in Table 2 and Table 3. As an additional measure of similarity, the Pearson correlation coefficient (PCC) was determined for each endogenous/synthetic peptide spectrum pair. Calculation of the PCC is sometimes limited to only those peaks that correspond to fragment ions predicted for the proposed sequence interpretation.^{28–30} Although this approach simplifies analysis, it introduces bias by excluding peaks that cannot be explained by the chosen interpretation and that might suggest that another sequence more accurately describes the spectrum. To avoid such bias, all peaks with intensities above a noise threshold were included in the analysis. Mirror plots illustrating the degree of similarity between MS/MS spectra for synthetic peptide validation standards and the endogenous peptides identified in mouse (Figure 2, Figures S-3 and S-4) and human (Figure 3, Figure S-5) islets are provided. Figure 4 and Figure S-6 summarize the results of the PCC analyses for these comparisons. For accurate retention time comparisons, islet samples and synthetic peptide samples were spiked with a retention time calibration mixture and retention times for the different runs were aligned based on the elution of the calibrants. The absolute value of the difference between the adjusted retention time for the synthetic and endogenous peptide (delta RT) is reported in the table. As an example, the MS/MS spectrum obtained by fragmentation of the synthetic LGGPGAGSLQPL-EAE peptide matched the spectrum acquired from human islets with a high degree of confidence

(% SPI = 93.4, score = 25.9, PCC = 0.94) and the endogenous and synthetic peptide eluted at the same time chromatographically (delta RT = 0.04 min). Thus, we can be highly confident that the HIP sequence LGGPGAGSLQPL-EAE is present in human islets.

In summary, 12 HIP matches satisfying the first three rules were identified during analysis of the NOD mouse islet samples (Table 2). Of these, one HIP sequence, DLQTLAL-WSRM, which was previously identified,¹ was fully validated. Three others satisfied all the criteria that were assessed but need further assessment by comparison to synthetic peptides before being fully validated. Analysis of BALB/c mouse islet samples led to the identification of nine HIP matches satisfying the first three rules (Table S-1). Of these nine, two HIP sequences, DLQTL-WSRM and DLQTLAL-WSRM, were fully validated and four passed the first five rules but have not yet been fully evaluated using synthetic peptides. Twelve HIP matches passing the first three rules were identified in the analysis of the human islet samples (Table 3). Of these, the HIP sequence LGGPGAGSLQPL-EAE was fully validated and two other matches satisfied all criteria that were assessed, but complete assessment with synthetic peptide standards is still needed. For both mouse and human samples, some HIP matches failed validation by only a small margin. These matches may still merit consideration and could potentially be validated by optimizing collection parameters to obtain higher quality spectra.

DISCUSSION

We previously identified a novel class of post-translationally modified peptides—hybrid insulin peptides—in the beta cells of NOD mice.¹ As a critical next step, we have identified here with high confidence the presence of a HIP in the islets of two different human donors, providing a rationale for further investigation into the role of HIPs in human disease. CD4 T cell responses to HIPs seem to be a major contributor to the pathogenesis of autoimmune diabetes in the NOD mouse^{1,2,4} and potentially in human T1D.^{1,3} The identification of HIPs in islets from the non-autoimmune BALB/c mouse strain and islets from human donors without T1D indicates that the presence of HIPs in the islets is not sufficient to cause disease. However, differences in the immune response to HIPs could be responsible for the development of T1D in predisposed individuals. Such differences could relate, for example, to the efficiency with which HIPs are presented by various major histocompatibility complex (MHC) molecules,⁵ how effectively peripheral tolerance to these antigens is maintained, or the timing or extent of HIP formation.

The confident identification of HIPs in islets will be critical to the development of a better understanding of their biological and disease relevance. To set a precedent for rigor in this challenging new area of proteomics, we applied an extensive set of custom criteria for use in validating potential HIPs identified by mass spectrometric analysis of islet samples. These criteria provide a starting point for eventual standardization and wide adoption by the community. Although the Spectrum Mill software was used for much of the analysis in our work, similar analyses could be performed with most other commonly used software. In recent years, the identification of peptides generated by proteasome-mediated peptide splicing^{31,32} has become an area of great interest in the field of immunopeptidomics, and work has begun to develop the novel approaches needed to identify such peptides.

10,29,30,33,34 Although it is not yet known if HIPs are generated by proteasome-mediated splicing or by a different mechanism, confident identification of tandem mass spectra corresponding to either HIPs or spliced peptides presents the same challenges. Thus, the guidelines reported here for the validation of putative HIP matches are of value in both areas and could be applied to emerging techniques for identifying potential HIPs or spliced peptides. Automation of this or other validation schemes would facilitate wide-scale application to larger data sets and across experiments.

It should be noted that failure to meet all the criteria assessed in this study does not necessarily indicate that a match was incorrect. For example, the HIP sequence LGGGPGAGSLQPL-EAE was identified and fully validated in human islets. A truncated version of this sequence, LGGGPGAGSLQPL-E (no missed GluC cleavages), and an extended version, LGGGPGAGSLQPL-EAEDLQVGQVE (two missed GluC cleavages), were also identified. The sequence LGGGPGAGSLQPL-E did not fully validate because only one right peptide residue was present in the sequence. The presence of only a single amino acid on one side of a HIP does not invalidate a HIP match but rather makes it impossible to determine the protein from which that side originated. Additionally, a fragment mass does not provide structural information on the fragment and could potentially originate from an entirely different residue with an identical mass. Confidence in mass spectral analyses is gained through the identification of peptide sequences that can be matched to expected sequences, such as the ones present in protein databases.

In many such cases, use of different proteases during sample preparation could generate different cleavage products containing more residues contributed by the left and/or right peptide. Poor b/y-ion coverage could indicate an incorrect match or could indicate that, although the match is correct, the particular fragment ions were not efficiently generated. This could be addressed in various ways for specific HIPs, such as optimizing fragmentation energy. In some cases, such as for the sequence match LGGGPGAGSLQPL-EAEDLQVGQVE, the spectra for the endogenous (islet-derived) peptide and the synthetic HIP validation peptide were too dissimilar to pass our filters even though all other criteria were satisfied. It is possible, for example, that in some cases spectral quality for the endogenous peptide was poor due to low precursor abundance and that a targeted run would generate a higher quality spectrum that would satisfy the spectral similarity criterion. However, in the present study, to maintain consistency and objectivity, such optimization was not included in the results. In the case of the sequence match LGGGPGAGSLQPL-EAE, the detection of two additional GluC cleavage products of the same HIP greatly increases confidence that this match is correct, even though the other two matches did not fully validate. Future efforts to automate validation of HIP matches could account for such cases by employing a probability-based approach. For example, cumbersome metrics such as left and right peptide b/y-ion coverage could be expressed as a probabilistic assessment of confidence in the left and right sequences. Then probabilities based on individual criteria could be combined to generate a single, probability-based score for each match. Rather than assigning absolute veto power to one criterion, such an approach would enable a more balanced, holistic assessment of confidence in a particular match. Furthermore, generation of a single, comprehensive score would simplify interpretation and facilitate automation of workflows.

The use of proteases in bottom-up proteomics introduces the possibility of generating hybrid peptides during sample preparation.^{35,36} When a peptide bond is attacked by the catalytic site of a protease such as GluC or trypsin, two products are generated: the free C-terminal peptide fragment and an acyl-enzyme intermediate consisting of the enzyme bound via an ester bond to the C-terminus of the N-terminal peptide fragment. Usually, a water molecule hydrolyzes this ester bond, liberating the N-terminal peptide fragment from the enzyme. However, if this bond is substituted by the N-terminal amine group of another peptide, a new peptide is formed in a process known as transpeptidation.³⁷ The HIP sequence LGGPGAGSLQPLALE-EAE was identified as a potential match for a spectrum obtained from the analysis of human islets. Given that GluC cleaves peptides at the C-terminus of glutamic acid (E) residues, it is possible that this HIP sequence could be formed by GluC cleaving insulin C-peptide, generating an acyl-enzyme intermediate with the peptide fragment LGGPGAGSLQPLALE, and then the N-terminal amine of insulin C-peptide cleaving the ester bond to form the hybrid. Although the sequences of most of the identified HIPs preclude such a mechanism, a few of the HIPs presented here could possibly be such artifacts of sample preparation. Future efforts to identify these same HIPs using different endoproteases for digestion would help rule out this possibility.

The identification of HIPs in this work highlights the previously underappreciated complexity of the beta cell proteome and the need to improve current approaches for analyzing tandem mass spectrometry data in order to account for the possibility of HIPs when interpreting spectra. Extending our previous identification of HIP sequences in NOD beta cells, our current findings suggest that HIPs may be a diverse class of post-translationally modified peptides. Further studies could utilize broader and more complex methodology (e.g., reduction and alkylation of samples, orthogonal chromatographic separation, more extensive in silico HIP databases) to interrogate the HIP landscape more deeply. Digestion of samples with various proteases would enable the identification of different peptides describing the same HIP junction, increasing confidence in the identification of a particular HIP. It is hoped that the effort involved in this work to ensure rigor when identifying HIPs in mouse and human islets will provide a new standard of analysis that will facilitate the accurate identification of HIPs in future efforts.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work was supported by American Diabetes Association grant 1-15-ACE-14 (TD) and NIH R01DK081166 (KH).

REFERENCES

- (1). Delong T; Wiles TA; Baker RL; Bradley B; Barbour G; Reisdorph R; Armstrong M; Powell RL; Reisdorph N; Kumar N; Elso CM; DeNicola M; Bottino R; Powers AC; Harlan DM; Kent SC; Mannering SI; Haskins K. Pathogenic CD4 T cells in type 1 diabetes recognize epitopes formed by peptide fusion. *Science* 2016, 351 (6274), 711–4. [PubMed: 26912858]

- (2). Wiles TA; Delong T; Baker RL; Bradley B; Barbour G; Powell RL; Reisdorph N; Haskins K. An insulin-IAPP hybrid peptide is an endogenous antigen for CD4 T cells in the non-obese diabetic mouse. *J. Autoimmun.* 2017, 78, 11–18. [PubMed: 27802879]
- (3). Babon JA; DeNicola ME; Blodgett DM; Crevecoeur I; Buttrick TS; Maehr R; Bottino R; Naji A; Kaddis J; Elyaman W; James EA; Haliyur R; Brissova M; Overbergh L; Mathieu C; Delong T; Haskins K; Pugliese A; Campbell-Thompson M; Mathews C; Atkinson MA; Powers AC; Harlan DM; Kent SC Analysis of self-antigen specificity of islet-infiltrating T cells from human donors with type 1 diabetes. *Nat. Med.* 2016, 22 (12), 1482–1487. [PubMed: 27798614]
- (4). Baker RL; Jamison BL; Wiles TA; Lindsay RS; Barbour G; Bradley B; Delong T; Friedman RS; Nakayama M; Haskins K. CD4 T Cells Reactive to Hybrid Insulin Peptides Are Indicators of Disease Activity in the NOD Mouse. *Diabetes* 2018, 67 (9), 1836–1846. [PubMed: 29976617]
- (5). Ito Y; Ashenberg O; Pyrdol J; Luoma AM; Rozenblatt-Rosen O; Hofree M; Christian E; Ferrari de Andrade L; Tay RE; Teyton L; Regev A; Dougan SK; Wucherpfennig KW Rapid CLIP dissociation from MHC II promotes an unusual antigen presentation pathway in autoimmunity. *J. Exp. Med.* 2018, 215 (10), 2617–2635. [PubMed: 30185635]
- (6). Doyle HA; Mamula MJ Posttranslational modifications of self-antigens. *Ann. N. Y. Acad. Sci.* 2005, 1050, 1–9.
- (7). Dunne JL; Overbergh L; Purcell AW; Mathieu C. Posttranslational modifications of proteins in type 1 diabetes: the next step in finding the cure? *Diabetes* 2012, 61 (8), 1907–14. [PubMed: 22826307]
- (8). Faridi P; Purcell AW; Croft NP In Immunopeptidomics We Need a Sniper Instead of a Shotgun. *Proteomics* 2018, 18 (12), No. e1700464. [PubMed: 29377634]
- (9). Wan X; Zinselmeyer BH; Zakharov PN; Vomund AN; Taniguchi R; Santambrogio L; Anderson MS; Lichti CF; Unanue ER Pancreatic islets communicate with lymphoid tissues via exocytosis of insulin peptides. *Nature* 2018, 560 (7716), 107–111. [PubMed: 30022165]
- (10). Gonzalez-Duque S; Azoury ME; Colli ML; Afonso G; Turatsinze JV; Nigi L; Lalanne AI; Sebastiani G; Carre A; Pinto S; Culina S; Corcos N; Bugliani M; Marchetti P; Armanet M; Diedisheim M; Kyewski B; Steinmetz LM; Buus S; You S; Dubois-Laforgue D; Larger E; Beressi JP; Bruno G; Dotta F; Scharfmann R; Eizirik DL; Verdier Y; Vinh J; Mallone R. Conventional and Neo-Antigenic Peptides Presented by beta Cells Are Targeted by Circulating Naive CD8+ T Cells in Type 1 Diabetic and Healthy Donors. *Cell Metab.* 2018, 28, 1–15. [PubMed: 29972795]
- (11). Deutsch EW; Csordas A; Sun Z; Jarnuczak A; Perez-Riverol Y; Ternent T; Campbell DS; Bernal-Llinares M; Okuda S; Kawano S; Moritz RL; Carver JJ; Wang M; Ishihama Y; Bandeira N; Hermjakob H; Vizcaino JA The ProteomeXchange consortium in 2017: supporting the cultural change in proteomics public data deposition. *Nucleic Acids Res.* 2017, 45 (D1), D1100–D1106. [PubMed: 27924013]
- (12). Vizcaino JA; Csordas A; del-Toro N; Dianas JA; Griss J; Lavidas I; Mayer G; Perez-Riverol Y; Reisinger F; Ternent T; Xu QW; Wang R; Hermjakob H. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* 2016, 44 (D1), D447–56. [PubMed: 26527722]
- (13). Verchere CB; Paoletta M; Neerman-Arbez M; Rose K; Irminger JC; Gingerich RL; Kahn SE; Halban PA Des-(27–31)C-peptide. A novel secretory product of the rat pancreatic beta cell produced by truncation of proinsulin connecting peptide in secretory granules. *J. Biol. Chem.* 1996, 271 (44), 27475–81. [PubMed: 8910330]
- (14). Boonen K; Baggerman G; D’Hertog W; Husson SJ; Overbergh L; Mathieu C; Schoofs L. Neuropeptides of the islets of Langerhans: a peptidomics study. *Gen. Comp. Endocrinol.* 2007, 152 (2–3), 231–41. [PubMed: 17559849]
- (15). Stevens AL; Wishnok JS; Chai DH; Grodzinsky AJ; Tannenbaum SR A sodium dodecyl sulfate-polyacrylamide gel electrophoresis-liquid chromatography tandem mass spectrometry analysis of bovine cartilage tissue response to mechanical compression injury and the inflammatory cytokines tumor necrosis factor alpha and interleukin-1beta. *Arthritis Rheum.* 2008, 58 (2), 489–500. [PubMed: 18240213]
- (16). Ellis J; Del Castillo E; Montes Bayon M; Grimm R; Clark JF; Pyne-Geithman G; Wilbur S; Caruso JA A preliminary study of metalloproteins in CSF by CapLC-ICPMS and NanoLC-CHIP/ITMS. *J. Proteome Res.* 2008, 7 (9), 3747–54. [PubMed: 18662025]

- (17). Alp O; Merino EJ; Caruso JA Arsenic-induced protein phosphorylation changes in HeLa cells. *Anal. Bioanal. Chem.* 2010, 398 (5), 2099–107. [PubMed: 20803194]
- (18). Wong SY; Lee CC; Ashrafzadeh A; Junit SM; Abraham N; Hashim OH A High-Yield Two-Hour Protocol for Extraction of Human Hair Shaft Proteins. *PLoS One* 2016, 11 (10), No. e0164993. [PubMed: 27741315]
- (19). Elias JE; Gygi SP Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* 2007, 4 (3), 207–14. [PubMed: 17327847]
- (20). Jeong K; Kim S; Bandeira N. False discovery rates in spectral identification. *BMC Bioinf.* 2012, 13, S2.
- (21). Dallas-Pedretti A; McDuffie M; Haskins K. A diabetes-associated T-cell autoantigen maps to a telomeric locus on mouse chromosome 6. *Proc. Natl. Acad. Sci. U. S. A* 1995, 92 (5), 1386–90. [PubMed: 7877988]
- (22). Keane TM; Goodstadt L; Danecek P; White MA; Wong K; Yalcin B; Heger A; Agam A; Slater G; Goodson M; Furlotte NA; Eskin E; Nellaker C; Whitley H; Cleak J; Janowitz D; Hernandez-Pliego P; Edwards A; Belgard TG; Oliver PL; McIntyre RE; Bhomra A; Nicod J; Gan X; Yuan W; van der Weyden L; Steward CA; Bala S; Stalker J; Mott R; Durbin R; Jackson IJ; Czechanski A; Guerra-Assuncao JA; Donahue LR; Reinholdt LG; Payseur BA; Ponting CP; Birney E; Flint J; Adams DJ Mouse genomic variation and its effect on phenotypes and gene regulation. *Nature* 2011, 477 (7364), 289–94. [PubMed: 21921910]
- (23). Zolg DP; Wilhelm M; Yu P; Knaute T; Zerweck J; Wenschuh H; Reimer U; Schnatbaum K; Kuster B. PROCAL: A Set of 40 Peptide Standards for Retention Time Indexing, Column Performance Monitoring, and Collision Energy Calibration. *Proteomics* 2017, 17 (21), 1700263.
- (24). Escher C; Reiter L; MacLean B; Ossola R; Herzog F; Chilton J; MacCoss MJ; Rinner O. Using iRT, a normalized retention time for more targeted measurement of peptides. *Proteomics* 2012, 12 (8), 1111–21. [PubMed: 22577012]
- (25). Hamaguchi K; Gaskins HR; Leiter EH NIT-1, a pancreatic beta-cell line established from a transgenic NOD/Lt mouse. *Diabetes* 1991, 40 (7), 842–9. [PubMed: 1647994]
- (26). Wang H; Qian WJ; Mottaz HM; Clauss TR; Anderson DJ; Moore RJ; Camp DG; Khan AH; Sforza DM; Pallavicini M; Smith DJ; Smith RD Development and evaluation of a micro- and nanoscale proteomic sample preparation method. *J. Proteome Res.* 2005, 4 (6), 2397–403. [PubMed: 16335993]
- (27). Guan Z; Yates NA; Bakhtiar R. Detection and characterization of methionine oxidation in peptides by collision-induced dissociation and electron capture dissociation. *J. Am. Soc. Mass Spectrom.* 2003, 14 (6), 605–13. [PubMed: 12781462]
- (28). Falth M; Svensson M; Nilsson A; Skold K; Fenyo D; Andren PE Validation of endogenous peptide identifications using a database of tandem mass spectra. *J. Proteome Res.* 2008, 7 (7), 3049–53. [PubMed: 18549260]
- (29). Liepe J; Marino F; Sidney J; Jeko A; Bunting DE; Sette A; Kloetzel PM; Stumpf MP; Heck AJ; Mishto M. A large fraction of HLA class I ligands are proteasome-generated spliced peptides. *Science* 2016, 354 (6310), 354–358. [PubMed: 27846572]
- (30). Faridi P; Li C; Ramarathinam SH; Vivian JP; Illing PT; Mifsud NA; Ayala R; Song J; Gearing LJ; Hertzog PJ; Ternette N; Rossjohn J; Croft NP; Purcell AW A subset of HLA-I peptides are not genomically templated: Evidence for cis- and trans-spliced peptide ligands. *Sci. Immunol* 2018, 3 (28), eaar3947. [PubMed: 30315122]
- (31). Hanada K; Yewdell JW; Yang JC Immune recognition of a human renal cancer antigen through post-translational protein splicing. *Nature* 2004, 427 (6971), 252–6. [PubMed: 14724640]
- (32). Vigneron N; Stroobant V; Chapiro J; Ooms A; Degiovanni G; Morel S; van der Bruggen P; Boon T; Van den Eynde BJ An antigenic peptide produced by peptide splicing in the proteasome. *Science* 2004, 304 (5670), 587–90. [PubMed: 15001714]
- (33). Mylonas R; Beer I; Iseli C; Chong C; Pak HS; Gfeller D; Coukos G; Xenarios I; Muller M; Bassani-Sternberg M. Estimating the Contribution of Proteasomal Spliced Peptides to the HLA-I Ligandome. *Mol. Cell. Proteomics* 2018, 17 (12), 2346–2357.

- (34). Rolfs Z; Solntsev SK; Shortreed MR; Frey BL; Smith LM Global Identification of Post-Translationally Spliced Peptides with Neo-Fusion. *J. Proteome Res.* 2018, DOI: 10.1021/acs.jproteome.8b00651.
- (35). Fodor S; Zhang Z. Rearrangement of terminal amino acid residues in peptides by protease-catalyzed intramolecular transpeptidation. *Anal. Biochem.* 2006, 356 (2), 282–90. [PubMed: 16859627]
- (36). Schaefer H; Chamrad DC; Marcus K; Reidegeld KA; Bluggel M; Meyer HE Tryptic transpeptidation products observed in proteome analysis by liquid chromatography-tandem mass spectrometry. *Proteomics* 2005, 5 (4), 846–52. [PubMed: 15759313]
- (37). Vigneron N; Ferrari V; Stroobant V; Abi Habib J; Van den Eynde BJ Peptide splicing by the proteasome. *J. Biol. Chem.* 2017, 292 (51), 21170–21179. [PubMed: 29109146]

Scheme 1.**Hybrid Insulin Peptide (HIP) Formation^a**

^aRepresentation of HIP formation, in which a C-terminally truncated C-peptide fragment (left peptide) combines with a cleavage product of another beta cell protein (right peptide) via a peptide bond. The mechanism of HIP formation is unknown.

Scheme 2.***In Silico* Generation of a Hybrid Insulin Peptide Sequence Database^a**

^aDatabase containing hypothetical HIP sequences was constructed by generating the list of possible truncated C-peptide sequences and combining each of these one at a time with the sequences of natural cleavage products of various beta cell proteins.

Figure 1.

Summary of peptides identified in NOD mouse islets and human islets by mass spectrometry. Islets from (a) NOD mice or (b) non-diabetic human donors were processed and proteins were digested with either AspN or GluC, respectively. The resultant peptide mixtures were analyzed by LC-MS/MS. Data were searched against a standard proteome database then against a custom HIP database. Proteins were ranked by the number of distinct peptides identified for the protein and the top eight proteins are reported. HIPs constituted less than 2% of the total number of distinct peptides identified. MS/MS spectra matched to HIPs were later subjected to further analysis to assess the validity of each match. Numbers reported are combined totals across three independent experiments. For each NOD islet experiment, islets were harvested and pooled from a separate group of NOD mice, and each human islet sample was from a different non-diabetic donor.

Figure 2.

Putative HIPs identified by mass spectrometry in islets from NOD mice validate with synthetic peptide standards. Following identification of potential HIPs in NOD islets, the putative HIP sequences were synthesized and the synthetic peptides were analyzed by tandem MS. Mirror plots comparing the MS/MS spectra for the endogenous islet peptide (positive y -axis) and the synthetic peptide standard (reflected y -axis) for two representative HIPs are shown. Band y -ions (singly and doubly charged) are displayed in blue and red, respectively. A mass tolerance of ± 20 ppm was used for labeling of band y -ions. Percent scored peak intensity (%SPI), score, and Pearson Correlation Coefficient (PCC) are indicated for each comparison. The difference in chromatographic retention time (ΔRT) for the peptide identified in islets as the sequence DLQTLAL-EVE and the synthetic validation peptide was not determined (n.d.).

Figure 3.

Putative HIPs identified by mass spectrometry in human islets validate with synthetic peptide standards. Following identification of potential HIPs in human islets, the putative HIP sequences were synthesized and the synthetic peptides were analyzed by tandem MS. Mirror plots comparing the MS/MS spectra for the endogenous islet peptide (positive y -axis) and the synthetic peptide standard (reflected y -axis) for two representative HIPs are shown. b - and y -ions (singly and doubly charged) are displayed in blue and red, respectively. A mass tolerance of ± 20 ppm was used for labeling of b - and y -ions. The two peptides identified here represent different GluC cleavage products of the same HIP, with the second containing a missed cleavage. Percent scored peak intensity (%SPI), score, and Pearson Correlation Coefficient (PCC) are indicated for each comparison. For the shorter peptide, a retention time match between the endogenous and synthetic peptide was not confirmed.

Figure 4.

Pearson correlation coefficients (PCC) obtained by comparison of mass spectra of endogenous peptides and synthetic HIPs. Fragmentation spectra for putative HIPs identified in NOD and human islets were compared to spectra obtained by analysis of synthetic versions of the peptides. For each comparison, peaks with an abundance greater than three standard deviations above the average in either spectrum were included. Error bars indicate 95% confidence intervals. Values that surpassed the PCC validation threshold of 0.9 are shown in green and those that did not are shown in red.

Table 1.**Validation Rules for Putative HIP Identifications**

RULE 1:	Spectrum can not be confidently matched to any peptides in a traditional protein database.
RULE 2:	Spectrum matches with high confidence and little ambiguity to a peptide in the HIP database.
RULE 3:	Peptide match corresponds to the proteolytic digest performed in sample preparation.
RULE 4:	The left peptide and right peptide regions each contain at least two amino acid residues.
RULE 5:	The spectrum contains sufficient b-/y-ion coverage of both the left and right peptide region of the HIP.
RULE 6:	The fragmentation spectrum for the endogenous peptide is highly similar to the spectrum obtained upon fragmentation of a synthetic version of the proposed HIP.
RULE 7:	The endogenous peptide and the synthetic HIP reference peptide elute at the same time by chromatography.

Table 2.

Validation of Putative HIPs Identified in the Islets of NOD Mice^a

Sequence ^a			endogenous peptide										synthetic peptide					
			Rule 1 ^c			Rule 2 ^c			Rule 3 ^c		Rule 4 ^c			Rule 5 ^c		Rule 6 ^c		Rule 7 ^c
			%SPI	score	RI-R2	%SPI	score	RI-R2	AspN	L pep	R pep	# of residues	L pep	R pep	by-ion coverage	%SPI	score	PCC
EVEDPQVAQLGGFGAGDLQTLALEVAQQ																		
EVED-LFVNSMTFG	ChgA	1	59.5	7.6	1.0	61.7	14.3	6.7	Yes	4	10	3	3	3	56.0	18.3	0.58	n.d.
EVED-TFVRSSTNFQM	IAPP	3	55.1	5.8	0.7	82.1	14.8	9.0	Yes	4	11	3	5	5	67.1	17.5	0.92	0.57
D-LFVNSMTFG	ChgA	3	84.7	10.4	0.3	93.1	13.9	3.5	Yes	1	10	0	4	4	n.d.	n.d.	n.d.	n.d.
D-TFVRSSTNFQM	IAPP	2	65.1	5.7	0.6	76.9	11.9	6.2	Yes	1	11	0	4	4	n.d.	n.d.	n.d.	n.d.
DFOVAQLG-PFSE	SG1	1	56.8	5.8	1.9	78.8	16.0	10.1	Yes	8	4	7	1	1	n.d.	n.d.	n.d.	n.d.
DFOVAQLGG-EVEDPQVAQLGGFGAG	C-pep	3	55.8	6.0	0.3	89.0	18.9	5.5	Yes	11	19	8	8	8	n.d.	n.d.	n.d.	n.d.
DFOVAQLGGFGAGDL-EVE	C-pep	1	46.0	5.4	0.1	82.1	11.8	4.3	Yes	18	3	8	2	2	n.d.	n.d.	n.d.	n.d.
DLOTI-L-NSRM	ChgA	2	61.6	9.2	1.2	89.2	17.6	8.4	Yes	5	4	4	4	4	86.2	21.8	0.66	0.42
DLOTI-L-EVE	C-pep	3	78.3	10.9	0.3	100.0	15.2	3.8	Yes	7	3	5	3	5	91.0	22.5	0.97	n.d.
DLOTI-L-LEGE	ChgA	1	61.9	10.3	1.2	94.8	17.1	6.4	Yes	7	4	4	4	4	91.4	22.8	0.86	n.d.
DLOTI-L-NAAR	IAPP	3	55.8	6.6	0.7	92.0	20.1	13.5	Yes	7	4	7	4	7	88.1	22.6	0.89	0.15
DLOTI-L-NSRM	ChgA	3	58.7	6.4	0.1	94.2	18.5	12.1	Yes	7	4	6	3	3	88.7	23.3	0.94	0.18
EVEDPQVAQLGGFGAGDLQTLALEVAQQ																		

^aInsulin 2 C-peptide sequence is shown in white letters as a reference. ^bNumber of samples out of three total in which putative HIP was identified. ^cBest match is reported for the SwissProt database search (methionine oxidation considered as a variable modification) and the HIP database search. No digest was specified for either search. Thresholds: %SPI > 70%, score > 10, RI-R2 > 2.5. To satisfy rule 1, the best match from the SwissProt database search was required to fall below at least one threshold. The HIP match was required to surpass all three thresholds to satisfy Rule 2. ^dThe match was required to be a sequence predicted for an AspN digest. A maximum of two missed cleavages were allowed. Only those matches that satisfied this criterion are shown in the table. ^eMore than one residue required for each side of the proposed HIP. ^fTwo or more band/or y-ions corresponding to fragmentation within the left peptide region and two or more corresponding to fragmentation within the right peptide region were required. If more than one detected ion corresponded to fragmentation at the same peptide bond, only one was counted. Fragmentation of the peptide bond at the left peptide-right peptide junction was counted for both L pep and R pep coverage. ^gSpectral similarity was assessed using the Spectrum Mill Spectrum Matcher utility. %SPI > 80 and score > 20 were required. The Pearson Correlation Coefficient (PCC) was calculated using GraphPad Prism software; a PCC > 0.9 was required for validation. ^hAll samples and standards were spiked with a retention time calibration mixture containing 40 standard peptides and observed retention times for the islet samples and synthetic peptide runs were aligned based on the retention times of these standards. Delta retention time threshold was 0.5 min. L pep = left peptide. R pep = right peptide. SPI = scored peak intensity. RI-R2 = rank 1 minus rank 2 score. RT = retention time. C-pep = insulin C-peptide. Seg1 = secretogranin 1. ChgA = chromogranin A. IAPP = islet amyloid polypeptide. n.d. = not determined. Values that satisfied the corresponding validation thresholds are highlighted in green, and those that did not are highlighted in yellow. For matches that validated across all categories, the HIP sequence is highlighted in green.

Table 3.

Validation of Putative HIPs Identified in Human Islets^a

Sequence ^a	R pep	samples ^b	Rule 1 ^c				Rule 2 ^c				Rule 3 ^c				Rule 4 ^c				Rule 5 ^c				Rule 6 ^c				Rule 7 ^c ΔRT (min)
			%SPI	score	R1-R2	%SPI	score	R1-R2	GluC	# of residues	L pep	R pep	by-ion coverage	L pep	R pep	%SPI	score	spectral similarity	L pep	R pep	%SPI	score	PCC				
EAEDLVQVQVLEGGFGAGSLQPLAEGLSLQ																											
ELQV ^d -FEQGLE	ChgA	1	56.3	6.9	0.1	81.6	12.2	4.0	Yes			4	7	1	3	2	3	2	76.1	19.4	0.65	nd	nd	nd	nd	nd	
LGGSFGAGSLQ-IE	ChgA	1	69.0	7.0	0.8	86.1	13.0	4.6	Yes			11	2	3	2	3	2	90.1	24.9	0.96	nd	nd	nd	nd	nd		
LGGSFGAGSLQPL-E	?	1	84.9	10.1	0.4	96.0	15.3	3.6	Yes			13	1	6	1	3	5	3	93.4	25.9	0.94	0.04	0.04	0.04	0.04	0.04	
LGGSFGAGSLQPL-EAE	C-pep	2	46.9	8.1	0.8	100.0	17.0	4.5	Yes			13	3	5	3	3	5	3	93.4	25.9	0.94	0.04	0.04	0.04	0.04	0.04	
LGGSFGAGSLQPL-EAEELQVQVE	C-pep	1	50.0	3.3	1.1	71.6	11.1	7.0	Yes			13	11	3	6	13	11	3	75.3	18.9	0.69	0.09	0.09	0.09	0.09	0.09	
LGGSFGAGSLQPL-GIVE	A-chain	1	48.4	4.2	1.1	81.4	10.9	6.8	Yes			13	4	5	3	13	4	5	3	75.3	18.9	0.69	0.09	0.09	0.09	0.09	
LGGSFGAGSLQPL-IE	ChgA	1	80.3	9.9	0.7	90.9	14.8	2.8	Yes			13	2	6	2	13	2	6	85.0	26.3	0.94	nd	nd	nd	nd	nd	
LGGSFGAGSLQPLA-E	?	2	78.4	8.8	1.7	91.7	12.9	4.1	Yes			14	1	7	1	14	1	7	85.0	26.3	0.94	nd	nd	nd	nd	nd	
LGGSFGAGSLQPLAIE-E	?	1	56.0	6.8	0.5	93.8	16.8	6.1	Yes			16	1	6	1	16	1	6	85.0	26.3	0.94	nd	nd	nd	nd	nd	
LGGSFGAGSLQPLAIE-EAE	C-pep	2	37.6	3.6	1.4	83.7	15.1	10.5	Yes			16	3	7	3	16	3	7	78.3	22.9	0.56	0.22	0.22	0.22	0.22	0.22	
LGGSFGAGSLQPLAIE-GIVE	A-chain	1	49.4	8.1	0.5	71.0	10.7	2.6	Yes			16	4	5	4	16	4	5	85.7	22.1	0.42	nd	nd	nd	nd	nd	
G-GIVEQCTSSISLYLE	A-chain	2	71.7	9.3	1.0	95.5	15.8	6.5	Yes			1	17	0	4	1	17	0	4	85.7	22.1	0.42	nd	nd	nd	nd	nd
EAEDLVQVQVLEGGFGAGSLQPLAEGLSLQ																											

^aInsulin C-peptide sequence is shown in white letters as a reference. ^bNumber of samples out of three in which putative HIP was identified. ^cBest match is reported for the SwissProt database search (methionine oxidation considered as a variable modification) and the HIP database search. No digest was specified for either search. Thresholds: %SPI > 70%, score > 10, R1-R2 > 2.5. To satisfy rule 1, the best match from the SwissProt database search was required to fall below at least one threshold. The HIP match was required to surpass all three thresholds to satisfy Rule 2. ^dThe match was required to be a sequence predicted for a GluC digest. A maximum of two missed cleavages were allowed. Only those matches that satisfied this criterion are shown in the table. ^eMore than one residue required for each side of the proposed HIP. ^fTwo or more b- and/or y-ions corresponding to fragmentation within the left peptide region and two or more corresponding to fragmentation within the right peptide region were required. If more than one detected ion corresponded to fragmentation at the same peptide bond, only one was counted. Fragmentation of the peptide bond at the left peptide-right peptide junction was counted for both L pep and R pep coverage. ^gSpectral similarity was assessed using the Spectrum Mill Spectrum Matcher utility. %SPI > 80 and score > 20 were required. The Pearson Correlation Coefficient (PCC) was calculated using GraphPad Prism software; a PCC > 0.9 was required for validation. ^hAll samples and standards were spiked with a retention time calibration mixture containing 40 standard peptides and observed retention times for the islet samples and synthetic peptide runs were aligned based on the retention times of these standards. Delta retention time threshold was 0.5 min. L pep = left peptide, R pep = right peptide, SPI = scored peak intensity, R1-R2 = rank 1 minus rank 2 score, RT = retention time, A-chain = insulin A-chain, C-pep = insulin C-peptide, ChgA = chromatogram A, n.d. = not determined. Values that satisfied the corresponding validation thresholds are highlighted in green, and those that did not are highlighted in yellow. For matches that validated across all categories, the HIP sequence is highlighted in green.